EXCEPTIONAL REDUCTION OF THE PLASTID GENOME OF SAGUARO CACTUS (*CARNEGIEA GIGANTEA*, CACTACEAE): LOSS OF THE *NDH* GENE SUITE AND INVERTED REPEAT[1]

Michael J. Sanderson[2], Dario Copetti[3,4], Alberto Búrquez[5], Enriquena Bustamante[5], Joseph Charboneau[2], Luis Eguiarte[6], Sudhir Kumar[7], Hyun Oh Lee[8], Junki Lee[9], Michelle McMahon[3], Kelly Steele[10], Rod Wing[3,4], Tae-Jin Yang[10], Derrick Zwickl[2], and Martin F. Wojciechowski[11]

[2]Dept. of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721 USA; [3]School of Plant Sciences, University of Arizona, Tucson, AZ 85721; [4]International Rice Research Institute, Genetic Resource Center, Los Baños, Laguna, Philippines; [5]Instituto de Ecologia, Unidad Hermosillo, UNAM, Hermosillo, Sonora, Mexico; [6]Instituto de Ecologia, UNAM Mexico D.F., Mexico; [7]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, Pennsylvania 19122, USA; [8]Phyzen Genomics Institute, 501-1, Gwanak Century Tower, Gwanak-gu, Seoul, 151-836, Republic of Korea; [9]Department of Plant Science, Plant Genomics and Breeding Institute, and Research Institute of Agriculture and Life Sciences, College of Agriculture and Life Sciences, Seoul National University, Seoul, 151-921, Republic of Korea; [10]Faculty of Science and Mathematics, College of Letters and Sciences, Arizona State University, Mesa, Arizona 85212 USA; [11]School of Life Sciences, Arizona State University, Tempe, Arizona 85287, USA

[1]Manuscript received _____; revision accepted _____.

The authors thank the University of Arizona-Universidad Nacional Autónoma de México Consortium for Drylands Research; Arizona State University; and the "Next-Generation BioGreen21 Program for Agriculture &Technology Development (Project No. PJ01100801)", Rural Development Administration, Korea, for funding; and Tumamoc Hill Reserve for allowing access and permission to collect saguaro tissue. We thank R. Jansen, T. Ruhlman, the Associate Editor and two anonymous reviewers for discussion and comments.

- *Premise of the study:* Land-plant plastid genomes have only rarely undergone significant changes in gene content and order. Thus, discovery of additional examples adds power to tests for causes of such genome-scale structural changes.
- *Methods:* Using next generation sequence data, we assembled the plastid genome of saguaro cactus and probed the nuclear genome for transferred plastid genes and functionally related nuclear genes. We combined these results with available data across Cactaceae and seed plants more broadly to infer the history of gene loss and to assess the strength of phylogenetic association between gene loss and loss of the inverted repeat (IR).
- *Key results:* The saguaro plastid genome is the smallest known for an obligately photosynthetic angiosperm (~113 kb), having lost the IR and plastid *ndh* genes. This supports a statistically strong association across seed plants between loss of *ndh* genes and loss of the IR. Many non-plastid copies of plastid *ndh* genes were found in the nuclear genome, but none had intact reading frames; nor did three related nuclear-encoded subunits. However, nuclear *pgr5*, which functions in a partially redundant pathway, was intact.
- *Conclusions:* The existence of an alternative pathway redundant with the function of the plastid NDH complex may permit loss of the plastid *ndh* gene suite in photoautotrophs like saguaro. This may provide a recurring mechanism for overall plastid genome size reduction, especially in combination with loss of the IR.

Key words: inverted repeat; *ndh* genes; phylogenomics; saguaro; cacti

Cacti are conspicuous elements of the arid and semi-arid succulent biome in the New World (Hernández-Hernández et al., 2014). Recent studies have elucidated many details of the timing and tempo of their diversification (Arakaki et al., 2011; Hernández-Hernández et al., 2011, 2014), their role in ecological interactions of arid ecosystems (Fleming et al., 2001; Bustamante et al., 2010), and diversification of associated species, such as cactophilic insects (Pfeiler and Markow, 2011). Molecular phylogenetic work has narrowed their position within the angiosperm clade Caryophyllales to a diverse subclade, Portulacineae (Nyffeler and Eggli, 2010), which also includes the remnants of the former Portulacaceae and the Madagascan succulent radiation Didiereaceae (Nyffeler, 2007; Arakaki et al., 2011). Multi-gene taxon-rich molecular phylogenies within cacti have established the monophyly of and many detailed relationship within two diverse subclades, Cactoideae and Opuntioideae, imbedded within a paraphyletic assemblage of the tropical genus *Pereskia* (Edwards et al., 2005; Hernández-Hernández et al., 2011). Despite their extreme morphology (Mauseth, 2006) and fascinating ecophysiological specializations (Silvera et al., 2010), genomic resources in cacti are still limited (Christin et al., 2015; Yang et al., 2015), and no plastid genome is available in cacti or other Portulacineae.

The saguaro cactus, *Carnegiea gigantea* (Engelm.) Britton & Rose, is both an iconic representative of Cactaceae and an example of its extreme specialization as one of the largest of the "columnar" forms that have evolved convergently several times (Hernández-Hernández et al., 2011). More than a century of demographic and ecological research has established its role as a keystone species in the Sonoran Desert (Pierson and Turner, 1998; Fleming et al., 2001; Pierson et al., 2013; Drezner, 2014), and its impact on

humans has been equally significant (Yetman, 2007). Here we focus on the remarkable structure of the saguaro plastid genome and attempt to understand aspects of the origin of its unique combination of features. Although the size, gene content, and structural organization of the plastid genome of land plants is highly conserved, these have been significantly modified in a few clades. For example, plastid genomes have undergone many rounds of inversions and other rearrangements in some conifers and in parts or all of the angiosperm families Campanulaceae, Leguminosae, Geraniaceae and Oleaceae (Cai et al., 2008; reviewed in Weng et al., 2014). Plastid genomes have independently lost one copy of the large inverted repeat (IR: ~25 kb in size) in conifers, in a large clade of papilionoid legumes, some species of *Erodium* (Geraniaceae), and the holoparasites *Conopholis* and *Phelipanche* (Orobanchaceae: Ruhlman and Jansen, 2014), leading to considerable reduction in the overall size of their plastid genomes. Non-photosynthetic angiosperms have commonly lost large components of the ancestral plastid gene complement (Braukmann et al., 2013; Wicke et al., 2013; Ruhlman and Jansen, 2014; Schelkunov et al., 2015), resulting in a highly reduced genome size. Finally, a small but trophically diverse assortment of taxa have lost most or all of the suite of 11 functionally related *ndh* genes (Braukmann et al., 2009; Wicke et al., 2011; Iles et al., 2013; Peredo et al., 2013; Ruhlman et al., 2015). These include specialized parasites (e.g., *Epifagus*, *Cuscuta*), carnivorous plants (e.g., *Genlisea*), mycoheterotrophs (e.g., *Rhizanthella*), aquatic plants (e.g., *Najas*), and xerophytes (e.g., *Ephedra*, *Welwitschia*), and some taxa with less unusual life histories, such as conifers (see Appendix S1 in Supplemental Data with the online version of this article).

We will show that most of these unusual modes of structural evolution have shaped the highly reduced plastid genome of the saguaro cactus, especially loss of the *ndh* gene suite and the inverted repeat. Both of these factors have potentially significant functional and evolutionary consequences. The plastid-encoded *ndh* subunits (A-K) join with a number of nuclear-encoded subunits to form the NADH dehydrogenase-like complex (NDH) (name follows Shikanai, 2014) found in the thylakoid membranes of chloroplasts of most land plants, which functions in cyclic electron flow (CEF) around photosystem I (Burrows et al., 1998; Ifuku et al., 2011; Martin and Sabater, 2010; Peng et al., 2011; Ueda et al., 2012; Shikanai, 2014). CEF is necessary for maintenance of efficient photosynthesis (Burrows et al., 1998; Munekaga et al., 2004; Ueda et al., 2012), and can also be important for photoprotection under water stress, high light intensity and high temperature (Burrows et al., 1998; Horvath et al., 2000; Wang et al., 2006). Recent work has also elucidated a second pathway which functions in CEF, the antimycin A sensitive ("AA sensitive") pathway, which is not dependent on plastid *ndh* genes and may actually be more important under most conditions (Shikanai, 2014).

Although loss of one redundant copy of the IR may not have functional consequences by itself, comparative studies have suggested it is associated with various kinds of "destabilization" of the plastid genome in some clades, as in the legume "inverted repeat lacking clade" (IRLC: Wojciechowski et al., 2004; Cai et al., 2008; Wicke et al., 2011). There the IR loss preceded rearrangements, gene losses (though the *ndh* genes were not lost) and a proliferation of repetitive sequences in some taxa (Palmer et al., 1987; Magee et al., 2010). However, further plastid genome sampling in the IRLC suggests the correlation between loss of the IR and rearrangements is weak at best, and

some other mechanism may trigger extensive rearrangements in some taxa of the IRLC (Sabir et al., 2014). More generally, gene losses and the loss of a copy of the IR provide a route to reduction in the overall size of the genome (Wu et al., 2009; Wicke et al., 2013), as do changes in the boundaries of the IR copies, which are variable in some taxa (Goulding et al., 1996; Ruhlman and Jansen, 2014).

We combine genomic data from both the saguaro plastid and nuclear genomes, with comparative data sampled across seed plants to begin to address several interrelated questions. How correlated are the loss of the IR and *ndh* suite generally across seed plants, and does one kind of loss typically precede the other in taxa experiencing both? What is the fate of the deleted *ndh* genes, both structurally and functionally? Is the loss compensated for by the presence of functional copies in the nuclear or mitochondrial genomes or by another pathway? Does the specialized lifestyle associated with some of the taxa that have lost their NDH complex provide clues to the functional consequences of *ndh* gene loss? Finally, is the driving force in these cases selection for a smaller genome (Wu et al., 2009) or something more directly related to the plastid NDH complex itself.

Not all of these questions can be answered equally rigorously with one additional genome sequence, even by placing it in a broader phylogenetic framework. Much of this paper therefore focuses on two key issues: understanding the possible association of *ndh* gene loss with loss of the IR; and understanding the genomic fate of the lost *ndh* genes. Our draft genome assembly for saguaro, together with that for pine (Zimin et al., 2014), is the only genome sequence data for taxa exhibiting both losses, but they are complemented by transcriptomic data just released for several other taxa that have lost

ndh genes (Ruhlman et al., 2015). Specific hypotheses for the fate of missing *ndh* genes include that they have been transferred to the nucleus but retain their function there; that their function has been coopted by other biochemical pathways; or that they are simply dispensable, at least under environmental conditions where they have been lost. To distinguish among these, it is important, first, to reconstruct the phylogenetic history of gene content to localize losses and infer the sequence of events. Although a complete picture of this in the case of saguaro will require many more plastid assemblies across cacti, we can leverage available data from other phylogenetic surveys across Portulacineae together with our assembly as a reference. Second, the rampant, though usually nonfunctional, transfer of plastid genes to the nuclear genome seen in other plant taxa (Matsuo et al., 2005; Lloyd and Timmis, 2011; Michalovova et al., 2013), provides a key hypothesis about the fate of these deleted genes. We will show by investigation of a draft nuclear genome assembly of saguaro whether this transfer has been effective. We can also test whether nuclear encoded genes in the AA sensitive CEF pathway are present in saguaro, which could compensate for the loss of CEF associated with loss of NDH complex (Shikanai, 2014).

MATERIALS AND METHODS

Plant material, DNA extraction, library construction, sequencing, and read preprocessing— Fresh tissue was obtained from two individual living plants of *Carnegiea gigantea* (Engelm.) Britton & Rose (see Appendix 1 for voucher information). DNA was extracted using a modification of the CTAB procedure (Doyle and Doyle, 1987). We constructed three paired end (PE) libraries and one mate pair (MP) library for

one accession and a single PE library for the other (Appendix S2, S3). All libraries were sequenced with Illumina HiSeq technology. Bases with quality value below 20 were removed from all reads with PrinSeqLite (Schmeider and Edwards, 2011), and the remaining reads were error-corrected using Bless (Heo et al., 2014) with a *k*-mer value of 45.

Plastid genome assembly— The 180 Myr divergence between cacti and the phylogenetically closest plastid genome assemblies in Amaranthanceae or Polygonaceae within Caryophyllales (divergence estimates per Magallón et al., 2015) makes reference guided assembly problematic. Several approaches to de novo assembly of plastid genomes from high throughput sequence data have been proposed. We used a strategy of downsampling reads to preferentially assemble the high copy number plastid genome (Sloan et al. 2014; Kim et al., submitted). We pursued this approach after initial assembly runs using all reads that typically returned plastid genes with lower accuracies (as indicated by comparisons to other saguaro sequences in GenBank) than highly conserved low copy number nuclear genes, possibly due to miss-assembly caused by divergent nuclear copies of these plastid genes. We used three different de novo assembly pipelines, involving three core assemblers, supplemented with meta-assembler tools in one protocol (see Appendix S2 for description of three protocols). These three protocols yielded five different assemblies across the two saguaro accessions (Table 1).

To obtain a consensus assembly in the context of problematic tandem repeats, we started with the smallest assembly, from Protocol 2 (and accession SGP5), and inserted repeat copies in the appropriate arrays if and only if the other four assemblies agreed on their placement and number. In four regions this led to insertion of 1-2 repeats each. In

addition, three bases were scored as ambiguous in the Protocol 2 assembly but were called unambiguously in the other four assemblies. We therefore used the consensus base call. Note that we are bringing to bear evidence from the assembly obtained from a second plant (SGP3) to the assembly of SGP5.

***Assessment of quality of the genome assembly*—**To assess the quality of the assemblies, we first compared the five to each other via multiple whole genome alignment with MAUVE (default options, assuming colinearity; Darling et al., 2004), in Geneious 8.0.5 (Kearse et al., 2012). Second, all 180 million read pairs in library SGP5-PE2 (280 bp nominal insert size; Appendix S3) were aligned against the five assemblies using bowtie2 (Langmead and Salzberg, 2012), and the resulting reads visualized in Tablet (Milne et al., 2013). Third, a small set of six targeted plastid sequences were available for saguaro from GenBank (Appendix S4), to serve as benchmark sequences. These included both genic and intergenic regions. Finally, based on the assembly, we designed PCR primers to five regions and sequenced them directly for confirmation (Appendix S5). Primers were designed with Primer3 v. 2.3.4 (Untergasser et al., 2012) in Geneious.

***Annotation*—**A baseline gene annotation was obtained using DOGMA (Wyman et al., 2004). Protein coding genes were checked and refined using blastx (Altschul et al., 1997) with protein sets derived from the phylogenetically nearest complete plastid genomes in Caryophyllales, *Fagopyrum esculentum* subsp. *ancestrale* (GenBank NC_010776.1; Logacheva et al., 2008), *Beta vulgaris* (GenBank EF534108.1), and *Spinacia oleracea* (NC_002202.1; Schmitz-Linneweber et al., 2001). Translations of

putative protein-coding genes were examined in all six reading frames in Geneious, and annotations were adjusted upstream or downstream by as many as 5 amino acids to coincide with start and stop codons as indicated. If no such codons were present, the annotation provided by DOGMA was retained by default.

Large repeats (>100 bp) were identified using 'repeat-match' in the MUMmer package, and sets of small tandem repeats (>20nt) were identified with 'exact-tandems' (Kurtz et al., 2004). Annotations were visualized using OGDRAW (Lohse et al., 2013) with a custom configuration file to permit addition of the tandem repeat regions (Fig. 1).

To visualize changes in gene order and content, we compared the saguaro assembly to the canonical gene order exemplified by *Spinacia oleracea* (GenBank accession NC_002202.1) using a newly generated annotation inferred with DOGMA to be consistent. This was facilitated by removing tRNA genes and mapping protein and ribosomal RNA genes only (Fig. 2).

Phylogenetic analysis of plastid genes and phylogenetic relationships of saguaro—We downloaded the comprehensive sets of plastid genes obtained in the study of Arakaki et al. (2011) for 10 Portulacineae, including the cacti *Pereskia, Maihuenia, Blossfeldia, Weingartia, Opuntia, Pereskiopsis*, and the outgroups *Anredera*, *Portulacaria*, *Portulaca*, *Didierea*, plus *Mollugo* as an outgroup outside of Portulacineae. All loci from saguaro containing at least 240 bp were used as a query against these 11 taxa from GenBank, and hits with greater than 80% coverage were retained, aligned using MUSCLE 3.8 (Edgar, 2004), and visualized in Geneious to check alignments at the amino acid level. A total of 40 protein coding genes and the 16S and 23S ribosomal

genes were retained, having four or more taxa present. Note that all *ndh* genes except *ndhD* were excluded, since most are missing in the saguaro assembly. The gene *ndhB*, though present in saguaro, was excluded because it is highly truncated. A supermatrix was assembled and phylogenetic tree inferred with GARLI v. 2.1 (Zwickl, 2006), using a model having a common GTR+I+G model, with separate rate modifiers for each locus. Bootstrapping was implemented with 400 replicates.

We tested for differences in synonymous and nonsynonymous rates of substitution using relative rates tests between saguaro and its two closest relatives in the inferred species tree. For each of the 40 plastid protein coding genes, we used HYPHY version 2.2.4 (Kosakovsky Pond et al., 2005), based on the Muse and Gaut (1994) codon substitution model, to compare saguaro rates to *Weingartia* and then to *Blossfeldia*. In both cases *Portulaca* was used as the outgroup.

Interrogation of saguaro whole genome assembly for copies of missing plastid genes— We searched for non-plastid homologs of the plastid *ndh* genes found in other cacti but missing in the saguaro plastid genome among the scaffolds constructed in a draft saguaro whole genome assembly (Copetti et al., in prep.; see Appendix S2). We used blastn and tblastn with a complete set of 11 *ndh* genes from the cactus *Maihuenia poeppigii* as queries, keeping all hits to scaffolds with an E-value > 1e-5 and percent identity > 60%. This returned a pool of candidate nuclear copies of plastid *ndh* genes.

To test whether these were mitochondrial rather than nuclear in origin, we assembled contigs from the saguaro mitochondrial genome using four repeated runs of assembly Protocol 3 (Appendix S2) with 6 million read pairs from our Illumina libraries.

A few of the resulting contigs were plastid and were excluded. The remaining contigs were checked against the *Beta vulgaris subsp. maritima* mitochondrial protein sets (downloaded as a set via GenBank Bioproject accession PRJNA62897) using blastx, keeping only those with similarity at a stringent E-value of 1e-100. This formed a pool that was enriched for mitochondrial contigs. The PE library SGP5-PE2 (Appendix S3) was aligned to each of a sample of five of these putative mitochondrial scaffolds using bowtie2 to compare read depth to that for the scaffolds built from the whole genome assembly. Finally, each candidate nuclear copy of a plastid *ndh* gene, found on a scaffold from the whole genome assembly, was checked against the set of mitochondrial contigs using blastn.

More detailed analysis focused on just *ndhF*, a commonly used phylogenetic marker that is missing from saguaro. Putative nuclear homologs of a *Maihuenia ndhF* query (accession HQ620944.1) were retrieved from the nuclear genome assembly by blastn searches, keeping all scaffolds with hits > 500 nt. The structure of the scaffolds was explored by pairwise blastn searches of each against the set of annotated saguaro plastid genes, as well as the putative mitochondrial scaffolds. A large PE library was aligned to each scaffold as above to examine read coverage. To validate the quality of the assembly of these non-plastid *ndhF* copies, in the context of potential problems induced by having numerous paralogous copies, we checked the assemblies against an independent assembly protocol (SOAPdenovo2 with just the PE libraries and default parameters) and compared sequence homology by reconstructing a maximum likelihood phylogenetic tree in RAxML 7.2.7 (Stamatakis, 2006) with the default GTRCAT model using scaffolds from both assembly protocols. We also designed primers to PCR amplify

and sequence two of these copies, together with their flanking regions for validation of the assembly.

Sequence from just the *ndhF* gene was extracted from nuclear scaffolds and added to a data set of 94 other plastid *ndhF* sequences sampled across Portulacineae, obtained from GenBank. Multiple sequence alignment was done with MUSCLE 3.8 (Edgar, 2004) with default parameters, visualized in Geneious for inspection, and a phylogenetic tree constructed with RAxML as above, with the fast bootstrap option.

We also looked for three nuclear-encoded subunits of the plastid NDH complex, *ndhM*, *ndhN*, and *ndhO* by using *Arabidopsis thaliana* genes in tblastn searches against the whole genome assembly. Similarly, we searched these nuclear scaffolds for two genes, *pgr5* and *pgrL1*, that encode key proteins in the AA sensitive CEF pathway and can be used to infer its presence (Ruhlman et al., 2015).

***Comparative method analysis of the IR loss and* ndh *gene suite loss*—**To test whether there is a phylogenetic association between loss of the IR and loss of the *ndh* gene suite broadly across seed plants, we used BayesTraitsV2 (Pagel, 1994; Pagel et al., 2004) in the context of a time-calibrated phylogeny of 798 seed plants (Magallón et al., 2015: downloaded from the Dryad data repository). Data on IR and *ndh* gene loss were taken from the literature and NCBI, and scored as a single binary trait for each. Data from complete plastid genome sequences (Appendix S1) was supplemented by three other strongly supported putative *ndh* gene losses in Santalales (Moore et al., 2010), Alismatales (Iles et al., 2013) and Ericaceae (Braukmann and Stefanovic, 2012; Appendix S6). Coding was approximate in two respects. First, since there is not exact taxon name matching between the tree and available genome data, we used available

phylogenetic trees from the literature to map these as closely as possible (Appendix S6). Second, a diverse clade having taxa with both states of character, e.g., Orobanchaceae, can be represented by a single leaf in the Magallón et al tree, (*Pedicularis*)--that is, be polymorphic. For both the IR and *ndh* characters, we coded the derived "1" state (losses) in the clade if *any* taxa exhibited it. We performed a likelihood ratio and AIC test of dependent and independent discrete models in the program, and a Bayes factor comparison of the two models was also undertaken with the MCMC option.

RESULTS

***Sequencing and assembly*—**Sequencing of PE and MP libraries for the two accessions (Table S3) generated 738 million read pairs, of which a small fraction was used in the plastid assemblies (Table 1). The five assemblies ranged in length from 112,872 to 113,300 nucleotides. A MAUVE multiple sequence alignment indicated 100% identity across all the genome assemblies for almost their full lengths, with the length differences being attributable to a few locations where small tandem repeat arrays have different copy numbers in the various assemblies (see Appendix S7 for visual comparison of assemblies). Our final consensus *in silico* assembly had a length of 113,064 bp and a G+C content of 36.7%. Alignment of the PE reads from library SGP5-PE2 indicated consistent coverage in sliding windows across the genome assemblies of ~8000x, reflecting the high copy number of the plastid genome in the isolated saguaro tissue.

Agreement of the consensus assembly to five plastid genic and intergenic sequences from saguaro obtained from GenBank was high (Appendix S4**;** see also

discussion in Appendix S2**)**, and Sanger sequencing of two of these confirmed 100% identity to the assembly.

***Size and structure of the saguaro plastid genome*—**An annotated map of the saguaro plastid genome is shown in Figure 1, indicating 71 protein coding genes, 4 rRNAs and 29 unique tRNAs (see Appendix S2 and S8 for annotation details). Several of the protein coding genes, including *rpl23*, *accD*, *ndhB* and *ndhD* are probably pseudogenes based on their structure. The size of the saguaro plastid genome is substantially smaller than other photosynthetic angiosperms (Appendix S9). Sanger sequencing of one of the regions of tandem repeats indicated some variation in repeat number between two accessions of saguaro, but since the range of assembled genome sizes does not exceed 113,300 bp across five different assemblies with much variation in inferred repeat structure (Table 1), it is unlikely that the genome size deviates much from our consensus assembly, and we therefore use the *in silico* assembly as a benchmark for saguaro's "reference" genome size. The overall reduction in size is due to a number of factors, most notably the loss of the canonical inverted repeat structure present ancestrally in angiosperms (Fig. 2).

Saguaro appears to have lost most of IRa combined with a large inversion involving IRb and the Small Single Copy (SSC) region (Fig. 2: region IV). Together this has served to place *ycf2* adjacent to *ycf1* on one side of the (now inverted) set of genes from the SSC region. On the other side are the ribosomal RNA genes that are normally between the two *ycf* genes (these second copies now missing).The boundaries of this

putative inversion event were confirmed experimentally with PCR and sequencing using primer pairs spanning the putative breakpoints (Fig. 2; Appendix S5).

Gene order in the ancestral large single copy region has been retained unchanged from the canonical order across angiosperms except for a small ~6 kb inversion involving four genes: *rbcL - atpB - atpE - trnM* (Fig. 2: region II) flanked by two partial copies of *trnF*.

The second notable reduction in the genome is the deletion or pseudogenization of the entire suite of 11 *ndh* genes typically found in plastid genomes. There are remnants of only two plastid *ndh* genes in the saguaro plastid assembly: *ndhB* and *ndhD*. Of the two, *ndhD* is full length with 501 codons, but it contains four stop codons. The *ndhB* gene is found in two exons, as is typical, but is missing about 200 residues in the middle of its coding region.

Additional features, including the loss of several introns and a novel structure of the *rps12* gene, are described in Appendices S2 and S8.

Phylogenetic position of saguaro and rates of plastid gene evolution— The maximum likelihood phylogeny based on 42 concatenated plastid genes (Fig. 3) strongly supports the placement of saguaro as sister to *Weingartia*, both of which are within the "core Cactoideae" clade recognized by Hernández-Hernández et al. (2011; see also Ritz et al., 2007). Analyses of relative rates of gene evolution uncovered only one notable result: the synonymous rate for *rbcL* in saguaro was significantly faster than *Weingartia* (or *Blossfeldia*) (see Appendix S2 for further discussion).

***Phylogenomics of* ndh *genes*—**A large number of copies of the 11 *ndh* genes in the plastid genome of the cactus, *Maihuenia*, were found in BLAST searches of the saguaro draft nuclear genome (Appendix S10). There were tblastn hits to 99 scaffolds; blastn hits to 46. Of these, 85 and 36 scaffolds respectively had no trace of homology to our collection of putative mitochondrial genome scaffolds, making them strong candidates to be regions containing true nuclear copies. Several scaffolds included (parts of) more than one *ndh* subunit, often reflecting sets of co-located *ndh* genes within their original operons (Fig. 2; Appendix S10). None of the gene hits were full length intact reading frames.

Lengths of different *ndh* genes are quite variable (as measured in spinach; Appendix S10), but there was a strong positive correlation between this length and the number of whole genome scaffolds containing BLAST hits to *Maihuenia ndh* genes (Appendix S10; $r = 0.86$, $P < 0.001$). This might be expected if copying and transfer of plastid genes occurs uniformly and randomly across the plastid genome: larger genes should be "hit" by copy events more frequently and these copies will retain some similarity detectable in BLAST searches (Matsuo et al., 2005). This holds even though average size should also decrease as a function of age of the copy (Michalovova et al., 2013), assuming rate of copying has also been approximately independent of age.

Detailed investigation of ndhF. Five scaffolds containing hits in blastn searches against *Maihuenia* plastid *ndhF* satisfied the search criteria. The structure of these scaffolds was quite variable (Appendix S11), with some having no other copies of saguaro plastid genes, and others having several. One scaffold (21894_cov70) contained a small region homologous to the putative mitochondrial genome, as well as other plastid

sequences. Another scaffold (17720) contained two divergent copies of *ndhF* separated by 31 kb (Appendix S11). A parallel SOAPdenovo2 assembly recovered the same top five scaffolds, and the sequences were almost identical between assemblies. In fact, sequence divergence was much lower between assemblies than between scaffolds (Appendix S12). For further validation of the scaffold sequences, we directly sequenced two of these copies together with extensive flanking sequence (~1.8 kb of scaffold 5683 and~1.9kb of scaffold 17720), and these differed from the assembly by only one bp in one scaffold.

Alignment of one PE library (Fig. 4A-E) against each of these five scaffolds indicated modest read depth of 62x-83x of reads to the three *ndhF* scaffolds having no or almost no other plastid genes (scaffolds 3896, 5683, and 17720; see Appendix 11), which is consistent with read depth seen in other demonstrably nuclear scaffolds in the whole genome assembly, suggesting these are indeed nuclear sequences. The two other scaffolds have average read depths of 361x and 2380x, reflecting large numbers of hits to the plastid gene fragments co-located on these scaffolds. These were typically locations where homologs of plastid genes are still present in the saguaro plastid genome. Interestingly, one of these two (scaffold 21894_cov70) has the "nuclear" read depth of ~50x across areas of the scaffold without plastid homologs.

By comparison, alignment of the same read library to a sample of five scaffolds constructed by stringent homology to mitochondrial loci, resulted in read depths of 489x-609x. Three scaffolds with lowest coverage mentioned above have much lower coverage than this and a fourth is somewhat lower, which makes it likely that three, if not four, of

these scaffolds are nuclear and not mitochondrial. Scaffold 13_cov74 is embedded in a mosaic of regions at different read depths, requiring further investigation.

Phylogenetic reconstruction placed all six *ndhF* copies (including two from one scaffold) on very long branches nested within a set of 94 plastid *ndhF* genes from Cactaceae (Fig. 5). Bootstrap support for their precise position is weak but it is clear that all copies originated at or after the crown group radiation of cacti, which is supported at 90% bootstrap level), and there appear to be two separate clades of *ndhF* genes.

BLAST searches against the whole genome assembly for three nuclear-encoded subunits of the NDH-like complex returned hits with truncations and/or stop codons in each case. A search for the nuclear *pgr5* gene, a member of the AA sensitive CEF pathway, on the other hand, found a full length intact copy with one intron. One other hit to a paralogous pseudogene was also found. However, BLAST searches for the *pgr*L1 gene thought to be part of a functional complex with PGR5 in that CEF pathway, failed to find any hits.

***Correlation between IR loss and* ndh *gene loss*—**A strong association was detected between loss of the IR and loss of *ndh* genes across 798 species of seed plants in the tree from Magallón et al. (2015). (BayesTraits analysis: -2log LR = 39.37; $P < 0.001$; 4 df.; ΔAIC = 31.37). Bayesian results were sensitive to the rate priors, but when priors were guided by the ML estimates of rates (e.g., uniform on [0,0.01]), they returned Bayes factors with comparable support levels. The significance of this association stems from repeated coincident origins of the two traits in several clades, which Maddison and Fitzjohn (2015) refer to as "replicated co-distribution", and therefore reflects stronger

evidence of association than a comparable *P*-value in traits originating only once (cf., their Fig. 1).

DISCUSSION

Unusual structure of the saguaro plastid genome— Saguaro has an unusual plastid genome organization, involving loss of the IR and deletion of almost all traces of the *ndh* gene suite. At ~113kb, it is the smallest known for an obligately photosynthetic angiosperm (Appendix S9). More broadly across seed plants, *Ephedra* has a slightly smaller plastid genome at 109kb, but its close relatives, *Gnetum* and *Welwitschia* are larger, though still less than 120kb. The Gnetales have lost the *ndh* gene suite but not the inverted repeat, so their reduction in size is due to streamlining of introns and intergenic regions and gene loss (Wu et al., 2009). The remaining plants with plastid genomes under 125kb have either lost the inverted repeat (IRLC legumes, and conifers), are plant parasites (*Cuscuta*), or mycoheterotrophs (orchids).

Figure 2 indicates the differences in gene order and gene content between saguaro and the canonical angiosperm plastid gene order. Gene order in the ancestral Large Single Copy (LSC) region is highly conserved except for a small inversion of four genes (region "II"), which was also inferred by Downie and Palmer (1994) by restriction site mapping. The remainder for the genome has been more highly disrupted by an inverted order for genes in the ancestral SSC region and loss of one copy of the IR. It is possible to explain these changes by postulating one very large inversion (Region IV), coupled with a large deletion of most of the gene content of IRa. This must be combined with five deletions, including three separate regions containing different *ndh* operons. The first two steps in

this scenario are very similar to a two-step scenario proposed by Wu et al. (2007) to explain the loss of the IR in *Pinus*, including an initial inversion covering a large portion of the IR and region. The conservation of gene order in the LSC suggests that the genome has certainly not been uniformly destabilized by the loss of the IR. Similarly in the legume IRLC, although some taxa having lost the IR have undergone a high level of rearrangements, most others have retained the canonical gene order in the LSC established early in papilionoid phylogenetic history with a large 50-kb inversion (Sabir et al., 2014).

One effect of these structural changes is to juxtapose the two large genes, *ycf1* and *ycf2*. The *ycf1* gene presents a puzzle: it is an open reading frame for its full length, but it is also littered with small tandem repeats of in-frame nucleotide sequence, which means there are stretches of homology to *ycf1* genes in other taxa, with intervening stretches of non-homology. This gene has been shown to function in transport of proteins into the plastid (Kikuchi et al., 2013). Saguaro's plastid genome has many small tandem repeat regions, some of which are correlated with inversion endpoints (Weng et al., 2014), including the endpoints of the small region II inversion and the interior of the *ycf1*, which is near the large region IV inversion endpoint (Fig. 2).

***The loss and fate of plastid* ndh *genes in saguaro*—**Saguaro's plastid genome is completely missing nine of 11 plastid *ndh* genes. The remaining genes, *ndhB* and *ndhD* are likely nonfunctional, as the first is missing sequence coding for ~200 amino acids, and the second has four internal stop codons. Saguaro is the only member of Portulacineae for which we now have confirmatory genomic evidence of absence of the other nine genes. The comprehensive survey of Portulacineae plastid genes by Arakaki et

al. (2011) found these genes in outgroups of cacti and for the cacti *Maihuenia* and *Pereskiopsis* but only reported 1-3 of the plastid *ndh* genes to be present in the Cactoideae taxa *Blossfeldia* and *Weingartia*. Assuming these absences in the database can be taken as is, they imply the loss of most *ndh* genes at around the origin of the Cactoideae. This must be verified with additional plastid genome reconstructions but is certainly suggestive that loss of the *ndh* genes may coincide with the most diverse radiation within the cacti.

As the *ndh* genes are ancestrally organized in discontiguous blocks (operons) around the plastid genome, multiple deletions must have occurred to explain their absence. We leveraged results from whole genome sequencing in saguaro (Copetti et al., in prep.) to test several hypotheses about the fate of these genes and their function. One hypothesis is that they have been transferred to the nucleus where they retain their function. Clearly, a large number of copies of all the *ndh* genes are present in the nuclear genome. BLAST searches retrieved over 100 scaffolds containing one or more *ndh* genes, and most of these scaffolds were likely of nuclear origin. However, none of these copies have intact reading frames of the expected length, so they are likely nonfunctional. A smaller number of scaffolds had some sequence similarity to putative mitochondrial contigs. Some of the latter are due to homology with mitochondrial NADH genes, but other mitochondrial elements were also the cause. The structure of five scaffolds containing hits to *ndhF* is described in Appendix S11 and shows the tremendous variability in these putative nuclear regions. We used the copy number differences between plastid, mitochondrial and nuclear sequences to conclude that at least 3-4 of these *ndhF*-containing scaffolds are true nuclear copies. This is in contrast to the findings

of Lin et al. (2015) who found that the *ndh* genes missing from orchid plastid genomes were concentrated in the mitochondrial genome (though again were non-functional). At least one *ndhF*-containing scaffold showed some evidence of relationship to the saguaro mitochondrial genome at least in places, but the read depth of the *ndhF* region proper of all five assemblies is much more in line with the low copy number nuclear genome than either organellar genome (Fig. 4).

Phylogenetic analysis of these six *ndhF*-like sequences provides some hints as to their antiquity and fate after transfer. A phylogenetic tree (Fig. 5) indicates an origin of all of them within Cactaceae rather than among more distant relatives in the much larger sample of Portulacineae that was included. Presumably this bounds their age to no more than the crown age of Cactaceae. Although the tree is not well enough resolved to indicate precisely where and when copy events occurred, there is strong bootstrap support for at least two clades of these genes (Fig. 5), suggesting at a minimum that there was more than one copy event that may then have duplicated within the nuclear genome. The fact that branch lengths are much longer for the nuclear copies than for the Cactaceae plastid *ndhF* genes is consistent with both the higher average substitution rates in the nuclear genome and a probable release from functional constraint if these sequences are all pseudogenized, as their structure implies. Some copies are phylogenetically closer to a fragment of *ndhF* for the Cactoideae species *Copiapoa* (Nyffeler, 2007: GenBank accession DQ855879.1), annotated as nonfunctional by that author. Further work will be required to confirm whether this is actually a plastid fragment of this gene, perhaps on its way to complete deletion, or if it is a nuclear copy amplified nonspecifically by PCR. The short branch length implies it is still evolving at plastid genome-like rates.

These findings are consistent with a picture emerging from studies of plastid gene transfer in model plants, where there is a high ongoing rate of transfer of segments of the plastid genome to the nuclear genome (Millen et al., 2001; Huang et al., 2005; Matsuo et al., 2005; Guo et al., 2008; Michalovova et al., 2013). In rice, for example, 0.2% of the nuclear genome is composed of plastid sequence (Matsuo et al., 2005), the equivalent of six entire plastid genomes. However, in general nuclear copies are nonfunctional and have a relatively brief half-life during which their components are shuffled, rearranged, deleted, and undergo rapid nucleotide sequence divergence. In rice, the largest copy, which is nearly full length, is found on chromosome 10, with very low sequence divergence from the ancestral plastid genome (Matsuo et al., 2005) indicating a recent origin. However, in soybean, sequence divergence is much higher (Michalovova et al., 2013), which is closer to what we observe among scaffolds containing *ndh* genes in saguaro. With increasing numbers of sequenced mitochondrial genes, it has also become more evident that transfers to the mitochondrial genome from the plastid genome are not infrequent, as are transfers from the mitochondrial to the nuclear genome (Adams and Palmer, 2003; Goremykin et al., 2009). In fact, by probing BAC libraries of known genome type, Lin et al. (2015) showed that *ndh* copies were present in the mitochondrial genome of some species that had lost their plastid copies.

Saguaro *ndh* gene transfers do not appear to maintain function. Ruhlman et al. (2015) analyzed transcriptomes of taxa with deleted plastid *ndh* genes in Pinaceae, Gnetales, Geraniaceae and Orchidaceae and found no functional copies of these genes, which is in line with our negative findings in surveying the saguaro nuclear genome for functional copies. They also did not find transcripts for nuclear encoded members of the

plastid NADH-like complex, consistent with our finding of pseudogenized sequences for *ndhM*, *ndhN*, and *ndhO*.

Perhaps the plastid NDH-like complex is simply dispensable (Martin and Sabater, 2010). There are negligible phenotypic effects of *ndh* mutants in *Marchantia* and tobacco under greenhouse/growth chamber conditions (Burrows et al., 1998; Ueda et al., 2012). However, saguaro is presumably under light, temperature, and water stress much of the time; temperature and water stress of cacti including the saguaro has been studied in both seedling and adult stages (e.g., Smith et al. 1984; Franco and Nobel, 1989). Conditions of environmental stress are those in which the functional consequences of *ndh* mutants have been shown to be significant (Burrows et al., 1998; Horvath et al., 2000; Wang et al., 2006).

This raises the possibility that the function of the NDH-dependent pathway for cyclic electron flow is fully assumed by another pathway, such as the AA-sensitive CEF pathway (e.g., Hertle et al., 2013). Recent evidence strongly suggests that both pathways use ferredoxin as an electron carrier in both linear electron flow and CEF during photosynthesis (Yamamoto et al., 2011; Leister and Shikanai, 2013; Shikanai, 2014). Mutants of the nuclear encoded gene, *pgr*5, which is thought to have an important role in the AA-sensitive CEF pathway, have seemingly more serious effects on growth and photosynthesis than mutants in the NDH-dependent pathway, indicating that the former pathway may be the more important of the two (Munekaga et al., 2004; Shikanai, 2014). We found a full length putatively functional copy of the *pgr5* gene in the saguaro whole genome assembly, providing some evidence that there remains in place a functional pathway for CEF in the saguaro. However, BLAST searches for another gene in the

pathway, for *pgrL1*, returned several neighboring small hits on a single scaffold each with only moderate identity, and several had stop codons. This gene may simply be missing from our initial whole genome assembly (see Appendix S2), as it always co-occured with *pgr5* in a survey of transcriptomes of taxa missing *ndh* genes (Ruhlman et al., 2015).

***Correlates and causes of IR and* ndh *gene losses in saguaro*—**Although we found a phylogenetically significant association between loss of the IR and loss of *ndh* genes across seed plants, the broad scope of this analysis necessitated some compromises. The mismatch between taxa in the phylogeny of Magallón et al. (2015) and taxa for which data on gene loss and IR loss were available, as well as the coding of polymorphic taxa as "loss", probably had relatively small impact on the results. Both kinds of losses are sufficiently rare but also sufficiently repeated in a very large phylogeny that their phylogenetic proximity to each other is quite improbable statistically, unless there is an association.

One explanation for the statistical association could be that loss of *ndh* genes is triggered by "destabilization" of the genome following loss of the IR, or some other genomic event. If so, one might IR loss to precedes loss of *ndh* genes, but the evidence is very mixed. In *Erodium*, loss of the IR precedes loss of the *ndh* genes, which characterizes the so-called "long-branch clade" (Blazier et al., 2011), but evidently the IR has been regained there (R. Jansen, pers. comm.). [In our trait coding, we regarded this as a "loss" of the IR, despite this apparent lability. See Appendix S6] In gymnosperms--assuming conifers are a clade with Gnetales outside of it--then the IR was lost at the base of conifers and *ndh* gene loss occurred later in Pinaceae. However, this scenario is

complicated by two issues: first, the uncertain position of Gnetales, which has also lost the *ndh* suite but not the IR, and is the sister group of Pinaceae in the gnepine hypothesis (Wang and Ran, 2014); second, the IR may have been independently lost in Pinaceae and Cuppressophytes (Wu et al., 2011). The IRLC in legumes (Wojciechowski et al., 2004) has not experienced any loss of *ndh* genes; and there are many losses of *ndh* genes listed in Appendix S1 in taxa that have not lost the IR, such as in Orchidaceae. Saguaro has lost both the IR and *ndh* genes, and it will be interesting to see which events came first by further phylogenetic analysis of plastid genome gene content in other Cactaceae. Curiously, there seem to be no cases in which *ndh* genes have been lost first and the IR later--but sample size is small.

A correlation between loss of *ndh* genes and the transition to a heterotrophic lifestyle, perhaps due to relaxed selection on photosynthetic pathways, has been noted by several authors (e.g. Braukmann and Stefanovic, 2012; Peredo et al., 2013; Wicke et al., 2013, 2014). However, there are clearly nuances to this. Not only are there several lineages of seemingly unexceptional autotrophs for which this explanation cannot hold--Pinaceae, Gnetales, a small clade of species within *Erodium* (Geraniaceae) and saguaro (Appendix S1)--but there are also lineages with more unusual trophic strategies, which are still at least partly autotrophic, including carnivorous (Wicke et al., 2014) and aquatic plants (Peredo et al., 2013), for which reductions in photosynthetic rates rather than complete loss of photosynthesis is the case.

An alternative explanation may be that the loss of *ndh* genes is simply a byproduct of wholesale streamlining of the plastid genome (Wu et al., 2009), which would explain the strong association across seed plants between loss of the IR and loss of

ndh genes but would not require either a mechanistic connection between the two or exact concordance of timing. Ultimately, to unravel the causes of loss of the *ndh* genes and IR in saguaro and other taxa, it may be necessary to combine whole genome data with ecological, biochemical and molecular studies in these clades. The diversity of plant trophic strategies and physiologies exhibited by the taxa that have undergone these structural modifications make universal explanations both particularly tantalizing and particularly unlikely to be true.

LITERATURE CITED

ADAMS, K. L., AND J. D. PALMER. 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Molecular Phylogenetics and Evolution* 29: 380-395.

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG, W. Q. MILLER, AND D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25: 3389-3402.

ARAKAKI, M., P. A. CHRISTIN, R. NYFFELER, A. LENDEL, U. EGGLI, R. M. OGBURN, E. SPRIGGS, et al. 2011. Contemporaneous and recent radiations of the world's major succulent plant lineages. *Proceedings of the National Academy of Sciences of the United States of America* 108: 8379-8384.

BLAZIER, J., M. M. GUISINGER, AND R. K. JANSEN. 2011. Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). *Plant Molecular Biology* 76: 263-272.

BRAUKMANN, T. W. A., M. KUZMINA, AND S. STEFANOVIC. 2009. Loss of all plastid *ndh* genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny. *Current Genetics* 55: 323-337.

BRAUKMANN, T., AND S. STEFANOVIC. 2012. Plastid genome evolution in mycoheterotrophic Ericaceae. *Plant Molecular Biology* 79: 5-20.

BRAUKMANN, T., M. KUZMINA, AND S. STEFANOVIC. 2013. Plastid genome evolution across the genus *Cuscuta* (Convolvulaceae): two clades within subgenus *Grammica* exhibit extensive gene loss. *Journal of Experimental Botany* 64: 977-989.

BURROWS, P. A., L. A. SAZANOV, Z. SVAB, P. MALIGA, AND P. J. NIXON. 1998. Identification of a functional respiratory complex in chloroplasts through analysis of tobacco mutants containing disrupted plastid ndh genes. *EMBO Journal* 17: 868-876.

BUSTAMANTE, E., A. CASAS, AND A. BURQUEZ. 2010. Geographic variation in reproductive success of *Stenocereus thurberi* (Cactaceae): Effects of pollination timing and pollinator guild. *American Journal of Botany* 97: 2020-2030.

CAI, Z. Q., M. GUISINGER, H. G. KIM, E. RUCK, J. C. BLAZIER, V. MCMURTRY, J. V. KUEHL, et al. 2008. Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. *Journal of Molecular Evolution* 67: 696-704.

CHRISTIN, P. A., M. ARAKAKI, C. P. OSBORNE, AND E. J. EDWARDS. 2015. Genetic enablers underlying the clustered evolutionary origins of C4 photosynthesis in angiosperms. *Molecular Biology and Evolution* 32: 846-858.

DARLING, A. C. E., B. MAU, F. R. BLATTNER, AND N. T. PERNA. 2004. MAUVE: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* 14: 1394-1403.

DOWNIE, S. R., AND J. D. PALMER. 1994. A chloroplast DNA phylogeny of the Caryophyllales based on structural and inverted repeat restriction site variation. *Systematic Botany* 19: 236-252.

DOYLE, J., AND J. DOYLE. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11-15.

DREZNER, T. D. 2014. The keystone saguaro (*Carnegiea gigantea*, Cactaceae): a review of its ecology, associations, reproduction, limits, and demographics. *Plant Ecology* 215: 581-595.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792-1797.

EDWARDS, E. J., R. NYFFELER, AND M. J. DONOGHUE. 2005. Basal cactus phylogeny: implications of *Pereskia* (Cactaceae) paraphyly for the transition to the cactus life form. *American Journal of Botany* 92: 1177–1188.

FLEMING, T. H., C. T. SAHLEY, J. N. HOLLAND, J. D. NASON, AND J. L. HAMRICK. 2001. Sonoran Desert columnar cacti and the evolution of generalized pollination systems. *Ecological Monographs* 71: 511-530.

FRANCO, A. C., AND P. S. NOBEL. 1989. Effect of nurse plants on the microhabitat and growth of cacti. *Journal of Ecology* 77: 870-886.

GOREMYKIN, V. V., F. SALAMINI, R. VELASCO, AND R. VIOLA. 2009. Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. *Molecular Biology and Evolution* 26: 99-110.

GOULDING, S. E., R. G. OLMSTEAD, C. W. MORDEN, AND K. H. WOLFE. 1996. Ebb and flow of the chloroplast inverted repeat. *Molecular & General Genetics* 252: 195-206.

GUO, X., S. RUAN, W. HU, D. CA, AND L. FAN. 2008. Chloroplast DNA insertions into the nuclear genome of rice: the genes, sites and ages of insertion involved. *Functional & Integrative Genomics* 8: 101-108.

HEO, Y., X. L. WU, D. M. CHEN, J. MA, AND W. M. HWU. 2014. BLESS: Bloom filter-based error correction solution for high-throughput sequencing reads. *Bioinformatics* 30: 1354-1362.

HERNANDEZ-HERNANDEZ, T., H. M. HERNANDEZ, J. A. DE-NOVA, R. PUENTE, L. E. EGUIARTE, AND S. MAGALLON. 2011. Phylogenetic relationships and evolution of growth form in Cactaceae (Caryophyllales, Eudicotyledoneae). *American Journal of Botany* 98: 44-61.

HERNANDEZ-HERNANDEZ, T., J. W. BROWN, B. O. SCHLUMPBERGER, L. E. EGUIARTE, AND S. MAGALLON. 2014. Beyond aridification: multiple explanations for the elevated diversification of cacti in the New World Succulent Biome. *New Phytologist* 202: 1382-1397.

HERTLE, A. P., T. BLUNDER, T. WUNDER, P. PESARESI, M. PRIBIL, U. ARMBRUSTER, AND D. LEISTER. 2013. PGRL1 is the elusive ferredoxin-plastoquinone reductase in photosynthetic cyclic electron flow. *Molecular Cell* 49: 511-523.

HORVATH, E. M., S. O. PETER, T. JOET, D. RUMEAU, L. COURNAC, G. V. HORVATH, T. A. KAVANAGH, et al. 2000. Targeted inactivation of the plastid *ndhB* gene in tobacco results in an enhanced sensitivity of photosynthesis to moderate stomatal closure. *Plant Physiology* 123: 1337-1349.

HUANG, C. Y., N. GRUNHEIT, N. AHMADINEJAD, J. N. TIMMIS, AND W. MARTIN. 2005. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiology* 138: 1723-1733.

IFUKU, K., T. ENDO, T. SHIKANAI, AND E. M. ARO. 2011. Structure of the chloroplast NADH dehydrogenase-like complex: nomenclature for nuclear-encoded subunits. *Plant and Cell Physiology* 52: 1560-1568.

ILES, W. J. D., S. Y. SMITH, AND S. W. GRAHAM. 2013. A well-supported phylogenetic framewrok for the monocot order Alismatales reveals multiple losses of the plastid NADH dehydrogenase complex and a strong long-branch effect. *In* P. Wilkin AND S. J. Mayo [eds.], Early Events in Monocot Evolution, 1-28. Cambridge University Press.

KEARSE, M., R. MOIR, A. WILSON, S. STONES-HAVAS, M. CHEUNG, S. STURROCK, S. BUXTON, et al. 2012. Geneious Basic: An integrated and extendable desktop software

platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647-1649.

KIKUCHI, S., J. BEDARD, M. HIRANO, Y. HIRABAYASHI, M. OISHI, M. IMAI, M. TAKASE, et al. 2013. Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science* 339: 571-574.

KIM, K., S. LEE, J. LEE, H. LEE, H. JOH, N. KIM, H. PARK, AND T. YANG. Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within *Panax ginseng* species. *PLoS One*, submitted ms.

KOSAKOVSKY POND, S. L., S. D. W. FROST, AND S. V. MUSE. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21: 676-679.

KURTZ, S., A. PHILLIPPY, A. L. DELCHER, M. SMOOT, M. SHUMWAY, C. ANTONESCU, AND S. L. SALZBERG. 2004. Versatile and open software for comparing large genomes. *Genome Biology* 5: 2, R12.

LANGMEAD, B., AND S. L. SALZBERG. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357-U354.

LEISTER, D., AND T. SHIKANAI. 2013. Complexities and protein complexes in the antimycin A-sensitive pathway of cyclic electron flow in plants. *Frontiers in Plant Science* 4.

LIN, C. S., J. J. W. CHEN, Y. T. HUANG, M. T. CHAN, H. DANIELL, W. J. CHANG, C. T. HSU, et al. 2015. The location and translocation of *ndh* genes of chloroplast origin in the Orchidaceae family. *Scientific Reports* 5.

LLOYD, A. H., AND J. N. TIMMIS. 2011. The origin and characterization of new nuclear genes originating from a cytoplasmic organellar genome. *Molecular Biology and Evolution* 28: 2019-2028.

LOGACHEVA, M. D., T. H. SAMIGULLIN, A. DHINGRA, AND A. A. PENIN. 2008. Comparative chloroplast genomics and phylogenetics of *Fagopyrum esculentum* ssp *ancestrale* - a wild ancestor of cultivated buckwheat. *BMC Plant Biology* 8.

LOHSE, M., O. DRECHSEL, S. KAHLAU, AND R. BOCK. 2013. OrganellarGenomeDRAW-a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research* 41: W575-W581.

MADDISON, W. P., AND R. G. FITZJOHN. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Systematic Biology* 64: 127-136.

MAGALLóN, S., S. GóMEZ-ACEVEDO, L. L. SáNCHEZ-REYES, AND T. HERNáNDEZ-HERNáNDEZ. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist* 10.1111/nph.13264 DOI: n/a-n/a.

MAGEE, A. M., S. ASPINALL, D. W. RICE, B. P. CUSACK, M. SEMON, A. S. PERRY, S. STEFANOVIC, et al. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Research* 20: 1700-1710.

MARTIN, M., AND B. SABATER. 2010. Plastid *ndh* genes in plant evolution. *Plant Physiology and Biochemistry* 48: 636-645.

MATSUO, M., Y. ITO, R. YAMAUCHI, AND J. OBOKATA. 2005. The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. *Plant Cell* 17: 665-675.

MAUSETH, J. D. 2006. Structure-function relationships in highly modified shoots of Cactaceae. *Annals of Botany* 98: 901-926.

MICHALOVOVA, M., B. VYSKOT, AND E. KEJNOVSKY. 2013. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species:

size, relative age and chromosomal localization. *Heredity* 111: 314-320.

MILLEN, R. S., R. G. OLMSTEAD, K. L. ADAMS, J. D. PALMER, N. T. LAO, L. HEGGIE, T. A. KAVANAGH, et al. 2001. Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. *Plant Cell* 13: 645-658.

MILNE, I., G. STEPHEN, M. BAYER, P. J. A. COCK, L. PRITCHARD, L. CARDLE, P. D. SHAW, AND D. MARSHALL. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* 14: 193-202.

MOORE, M. J., P. S. SOLTIS, C. D. BELL, J. G. BURLEIGH, AND D. E. SOLTIS. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences of the United States of America* 107: 4623-4628.

MUNEKAGA, Y., M. HASHIMOTO, C. MIYAKA, K. I. TOMIZAWA, T. ENDO, M. TASAKA, AND T. SHIKANAI. 2004. Cyclic electron flow around photosystem I is essential for photosynthesis. *Nature* 429: 579-582.

MUSE, S. V., AND B. S. GAUT. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* 11: 715-724.

NYFFELER, R. 2007. The closest relatives of cacti: Insights from phylogenetic analyses of chloroplast and mitochondrial sequences with special emphasis on relationships in the tribe Anacampseroteae. *American Journal of Botany* 94: 89-101.

NYFFELER, R., AND U. EGGLI. 2010. Disintegrating Portulacaceae: A new familial classification of the suborder Portulacineae (Caryophyllales) based on molecular and

morphological data. *Taxon* 59: 227-240.

PAGEL, M. 1994. Detecting correlated evolution on phylogenies - a general-method for the comparative-analysis of discrete characters. *Proceedings of the Royal Society B-Biological Sciences* 255: 37-45.

PAGEL, M., A. MEADE, AND D. BARKER. 2004. Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology* 53: 673-684.

PALMER, J. D., B. OSORIO, J. ALDRICH, AND W. F. THOMPSON. 1987. Chloroplast DNA evolution among legumes - loss of a large inverted repeat occurred prior to other sequence rearrangements. *Current Genetics* 11: 275-286.

PENG, L. W., H. YAMAMOTO, AND T. SHIKANAI. 2011. Structure and biogenesis of the chloroplast NAD(P)H dehydrogenase complex. *Biochimica et Biophysica Acta-Bioenergetics* 1807: 945-953.

PEREDO, E., U. KING, AND D. LES. 2013. The plastid genome of *Najas flexilis*: adaptation to submersed environments is accompanied by the complete loss of the *ndh* complex in an aquatic angiosperm. *PLoS One* 8: e68591.

PFEILER, E., AND T. A. MARKOW. 2011. Phylogeography of the cactophilic Drosophila and other arthropods associated with cactus necroses in the Sonoran Desert. *Insects* 2: 218-231.

PIERSON, E. A., AND R. M. TURNER. 1998. An 85-year study of saguaro (*Carnegiea gigantea*) demography. *Ecology* 79: 2676-2693.

PIERSON, E. A., R. M. TURNER, AND J. L. BETANCOURT. 2013. Regional demographic trends from long-term studies of saguaro (*Carnegiea gigantea*) across the northern Sonoran Desert. *Journal of Arid Environments* 88: 57-69.

RITZ, C. M., L. MARTINS, R. MECKLENBURG, V. GOREMYKIN, AND F. H. HELLWIG. 2007. The molecular phylogeny of *Rebutia* (Cactaceae) and its allies demonstrates the influence of paleogeography on the evolution of South American mountain cacti. *American Journal of Botany* 94: 1321-1332.

RUHLMAN, T. A., AND R. K. JANSEN. 2014. The plastid genomes of flowering plants. *In* P. Maliga [ed.], Chloroplast Biotechnology: Methods and Protocols, vol. 1132, Methods in Molecular Biology, 3-38.

RUHLMAN, T., W.-J. CHANG, J. C. CHEN, Y.-T. HUANG, M.-T. CHAN, J. ZHANG, D.-C. LIAO, et al. 2015. NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. *BMC Plant Biology* 15: 100.

SABIR, J., E. SCHWARZ, N. ELLISON, J. ZHANG, N. A. BAESHEN, M. MUTWAKIL, R. JANSEN, AND T. RUHLMAN. 2014. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. *Plant Biotechnology Journal* 12: 743-754.

SCHELKUNOV, M., V. SHTRATNIKOVA, M. NURALIEV, M.-A. SELOSSE, A. PENIN, AND M. LOGACHEVA. 2015. Exploring the limits for reduction of plastid genomes: A case study of the mycoheterotrophic orchids *Epipogium aphyllum* and *Epipogium roseum*. . *Genome Biology and Evolution* 7: 1179-1191.

SCHMIEDER, R., AND R. EDWARDS. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27: 863-864.

SCHMITZ-LINNEWEBER, C., R. M. MAIER, J. P. ALCARAZ, A. COTTET, R. G. HERRMANN, AND R. MACHE. 2001. The plastid chromosome of spinach (*Spinacia oleracea)*: complete nucleotide sequence and gene organization. *Plant Molecular Biology*

45: 307-315.

SHIKANAI, T. 2014. Central role of cyclic electron transport around photosystem I in the regulation of photosynthesis. *Current Opinion in Biotechnology* 26: 25-30.

SILVERA, K., K. M. NEUBIG, W. M. WHITTEN, N. H. WILLIAMS, K. WINTER, AND J. C. CUSHMAN. 2010. Evolution along the crassulacean acid metabolism continuum. *Functional Plant Biology* 37: 995-1010.

SLOAN, D. B., D. A. TRIANT, N. J. FORRESTER, L. M. BERGNER, M. WU, AND D. R. TAYLOR. 2014. A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). *Molecular Phylogenetics and Evolution* 72: 82-89.

SMITH, S. D., B. DIDDEN-SOPFY, AND P. S. NOBEL. 1984. High-temperature responses of North American cacti. *Ecology* 65: 643-651.

STAMATAKIS, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688-2690.

UEDA, M., T. KUNIYOSHI, H. YAMAMOTO, K. SUGIMOTO, K. ISHIZAKI, T. KOHCHI, Y. NISHIMURA, AND T. SHIKANAI. 2012. Composition and physiological function of the chloroplast NADH dehydrogenase-like complex in *Marchantia polymorpha*. *Plant Journal* 72: 683-693.

UNTERGASSER, A., I. CUTCUTACHE, T. KORESSAAR, J. YE, B. C. FAIRCLOTH, M. REMM, AND S. G. ROZEN. 2012. Primer3-new capabilities and interfaces. *Nucleic Acids Research* 40.

WANG, P., W. DUAN, A. TAKABAYASHI, T. ENDO, T. SHIKANAI, J. Y. YE, AND H. L. MI. 2006. Chloroplastic NAD(P)H dehydrogenase in tobacco leaves functions in

alleviation of oxidative damage caused by temperature stress. *Plant Physiology* 141: 465-474.

WANG, X.-Q., AND J.-H. RAN. 2014. Evolution and biogeography of gymnosperms. *Molecular Phylogenetics and Evolution* 75: 24-40.

WENG, M. L., J. C. BLAZIER, M. GOVINDU, AND R. K. JANSEN. 2014. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. *Molecular Biology and Evolution* 31: 645-659.

WICKE, S., B. SCHAFERHOFF, C. W. DEPAMPHILIS, AND K. F. MULLER. 2014. Disproportional plastome-wide increase of substitution rates and relaxed purifying selection in genes of carnivorous Lentibulariaceae. *Molecular Biology and Evolution* 31: 529-545.

WICKE, S., G. M. SCHNEEWEISS, C. W. DEPAMPHILIS, K. F. MULLER, AND D. QUANDT. 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Molecular Biology* 76: 273-297.

WICKE, S., K. F. MULLER, C. W. DE PAMPHILIS, D. QUANDT, N. J. WICKETT, Y. ZHANG, S. S. RENNER, AND G. M. SCHNEEWEISS. 2013. Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. *Plant Cell* 25: 3711-3725.

WOJCIECHOWSKI, M., M. LAVIN, AND M. SANDERSON. 2004. A phylogeny of legumes (Leguminosae) based on analyses of the plastid *matK* gene resolves many well-supported subclades within the family. *American Journal of Botany* 91: 1846-1862.

WU, C. S., Y. N. WANG, S. M. LIU, AND S. M. CHAW. 2007. Chloroplast genome (cpDNA)

of *Cycas taitungensis* and 56 cp protein-coding genes of *Gnetum parvifolium*: Insights into cpDNA evolution and phylogeny of extant seed plants. *Molecular Biology and Evolution* 24: 1366-1379.

WU, C.-S., Y.-T. LAI, C.-P. LIN, Y.-N. WANG, AND S.-M. CHAW. 2009. Evolution of reduced and compact chloroplast genomes (cpDNAs) in Gnetophytes: selection toward a lower-cost strategy. *Molecular Phylogenetics and Evolution* 52: 115-124.

WU, C. S., Y. N. WANG, C. Y. HSU, C. P. LIN, AND S. M. CHAW. 2011. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and Cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. *Genome Biology and Evolution* 3: 1284-1295.

WYMAN, S. K., R. K. JANSEN, AND J. L. BOORE. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20: 3252-3255.

YAMAMOTO, H., L. PENG, Y. FUKAO, AND T. SHIKANAI. 2011. An SRC homology 3 domain-like fold protein forms a ferredoxin binding site for the chloroplast NADH dehydrogenase-like complex in Arabidopsis. *Plant Cell* 23: 1480-1493.

YANG, Y., M. J. MOORE, S. F. BROCKINGTON, D. SOLTIS, G. K. WONG, E. J. CARPENTER, Y. ZHANG, et al. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Mol Biol Evol*.

YETMAN, D. 2007. The great cacti: ethnobotany and biogeography. University of Arizona Press, Tucson, Arizona, USA.

ZIMIN, A., K. A. STEVENS, M. CREPEAU, A. HOLTZ-MORRIS, M. KORIABINE, G. MARCAIS, D. PUIU, et al. 2014. Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics* 196: 875-890.

ZWICKL, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.. Ph.D. thesis, University of Texas at Austin, Austin, Texas, USA.

Table 1. Descriptive statistics on five assemblies

Assembly ID[a]	Assembler(s)	Plant accession	Total read pairs used (millions)	Ungapped assembly length (bp)	Gapped assembly length (bp)[b]
1	CLC Genome Assembler	SGP3	4	113,145	113,145
2	SOAPdenovo 2.04; Amos 3.1.0;Tigr Assembler 2.0	SGP5	140	112,872	112,872
3 (scaffold-19)	Ray 2.3.1	SGP5	6	113,189	113,802
4 (scaffold-2)	Ray 2.3.1	SGP5	6	113,204	113,941
5 (scaffold-36)	Ray 2.3.1	SGP5	6	113,300	114,661
6 (sgp5_cp_frz1)	Consensus assembly[c]	SGP5	-	113,064	-

[a]Scaffold identifiers refer to Ray assembler output filenames

[b]Gaps added by Ray assembler based on inferred insert sizes only

[c]See text for methodology

Appendix 1. Voucher specimen information

Taxon; GenBank accession number of genome sequence; Collection Number; *Voucher specimen*, Collection locale (GPS coordinates); Herbarium.

Carnegiea gigantea (Engelm.) Britton & Rose; ----; Sanderson SGP3; *XXXX*; USA, B&B Cactus Nursery, Tucson, AZ (----); ARIZ. ***Carnegiea gigantea*** (Engelm.) Britton & Rose; XXXX; Sanderson SGP5; *XXXX*; USA, Tumamoc Hill Reserve, Tucson, AZ (N32.22003 degs.; W111.00343 degs.); ARIZ.

---- = not applicable

XXXX=to be added in proof

Figure Legends

Fig. 1. Annotated saguaro (*Carnegiea gigantea*) plastid genome, rendered in OGDRAW (Lohse et al., 2013). Features labeled "TR*n* (*Y*x) are blocks of tandem repeats labeled with integers, *n*, and the number of units within each block are indicated by *Y*.

Fig. 2. Plastid genome structure and gene order in saguaro (*Carnegiea gigantea*) compared to spinach. Spinach has the canonical order typical of most angiosperms. For simplicity, the circular map has been linearized and tRNAs have been deleted. Black triangles represent blocks present in spinach but absent in saguaro. A dashed line between the two genomes is a single gene translocation. Colinear block regions labeled I-V are discussed in text. Regions II and IV are inversions. IRa and IRb refer to the inverted repeat copies present in spinach; LSC=large single copy region; SSC=small single copy region. Small black rectangles indicate locations of five primer pairs designed to amplify key regions to check assembly with Sanger sequencing.

Fig. 3. Phylogenetic tree of Portulacineae, including saguaro (*Carnegiea gigantea*). Tree is a maximum likelihood analysis of 42 concatenated plastid genes. Bootstrap values are shown next to each clade.

Fig. 4. Read coverage depth of paired-end read library aligned to various scaffolds. A-E. Whole genome assembly scaffolds containing copies of plastid *ndhF* gene. Position of each copy indicated by black arrow. In each case the position maps to areas of relatively low coverage (~ 62-83x), even when surrounded by sequence with much higher read depth from copies of plastid genes still found in the plastid genome. Star indicates regions of ~1.8 kb that were directly sequenced for verification. Note, inset included in

panel A to show better the location of copy; peak coverage in panel E is truncated at 1000x for display. F. Coverage depth of a representative scaffold assembled using Protocol 3 in effort to identify set of putative mitochondrial scaffolds.

Fig. 5. ML tree of *ndhF* gene copies in whole genome scaffolds of saguaro and plastid copies of other Cactaceae. Figure shows just the Cactaceae clade within the entire Portulacineae analysis, with bootstrap values > 90% indicated by "*".

Figure 1

Click here to download Figure: Saguaro_cpFigs.revision 1.pdf

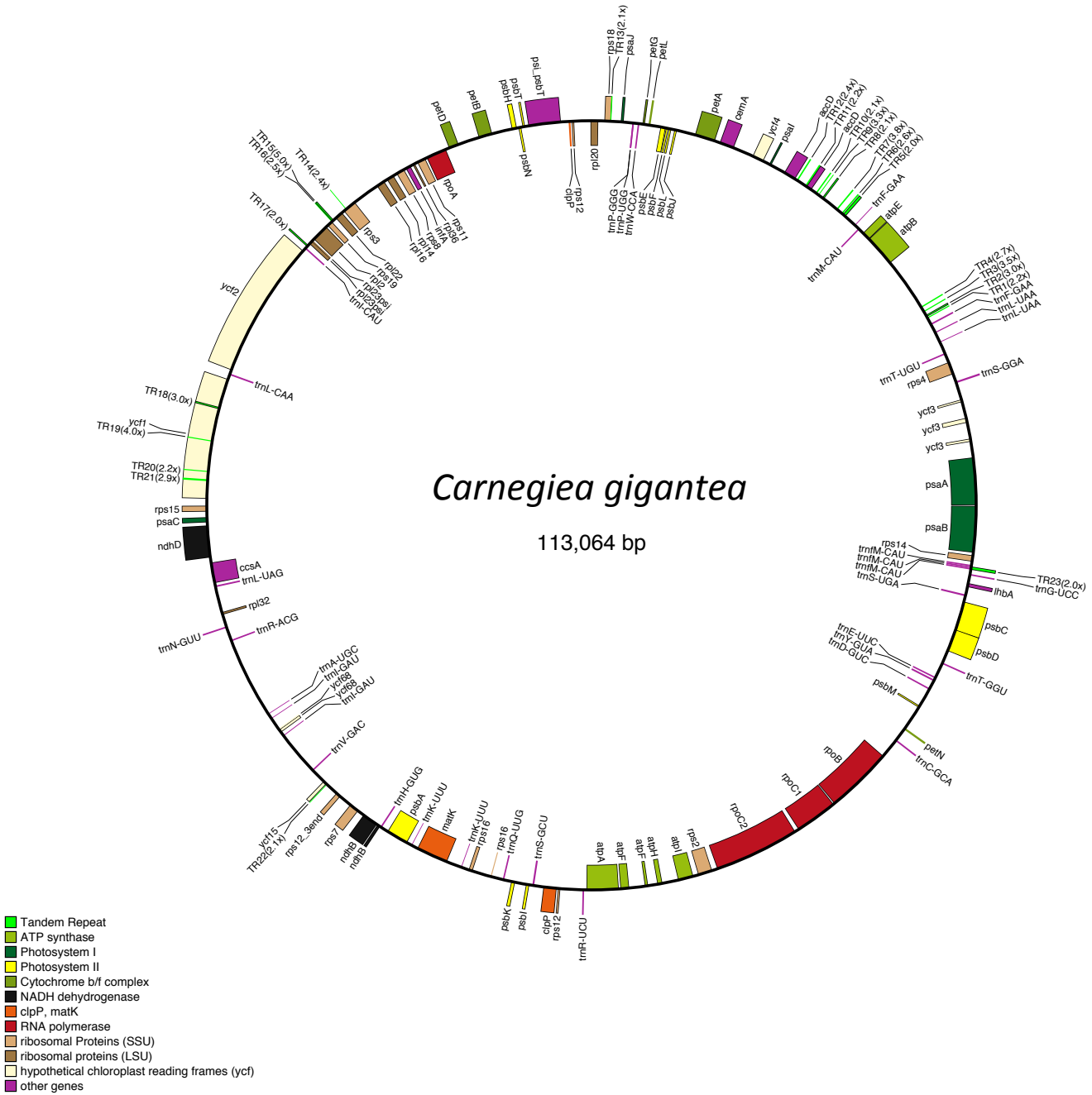

Figure 2

Click here to download Figure: Saguaro_cpFigs.revision 2.pdf

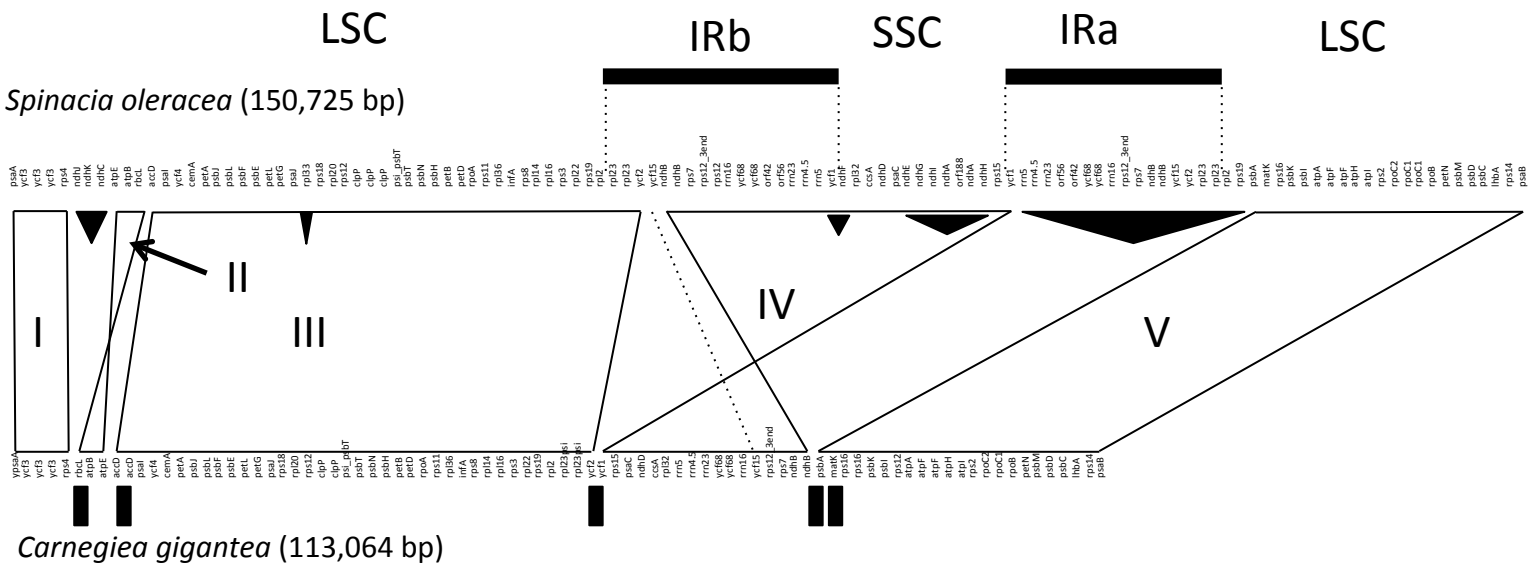

Figure 3

Click here to download Figure: Saguaro_cpFigs.revision 3.pdf

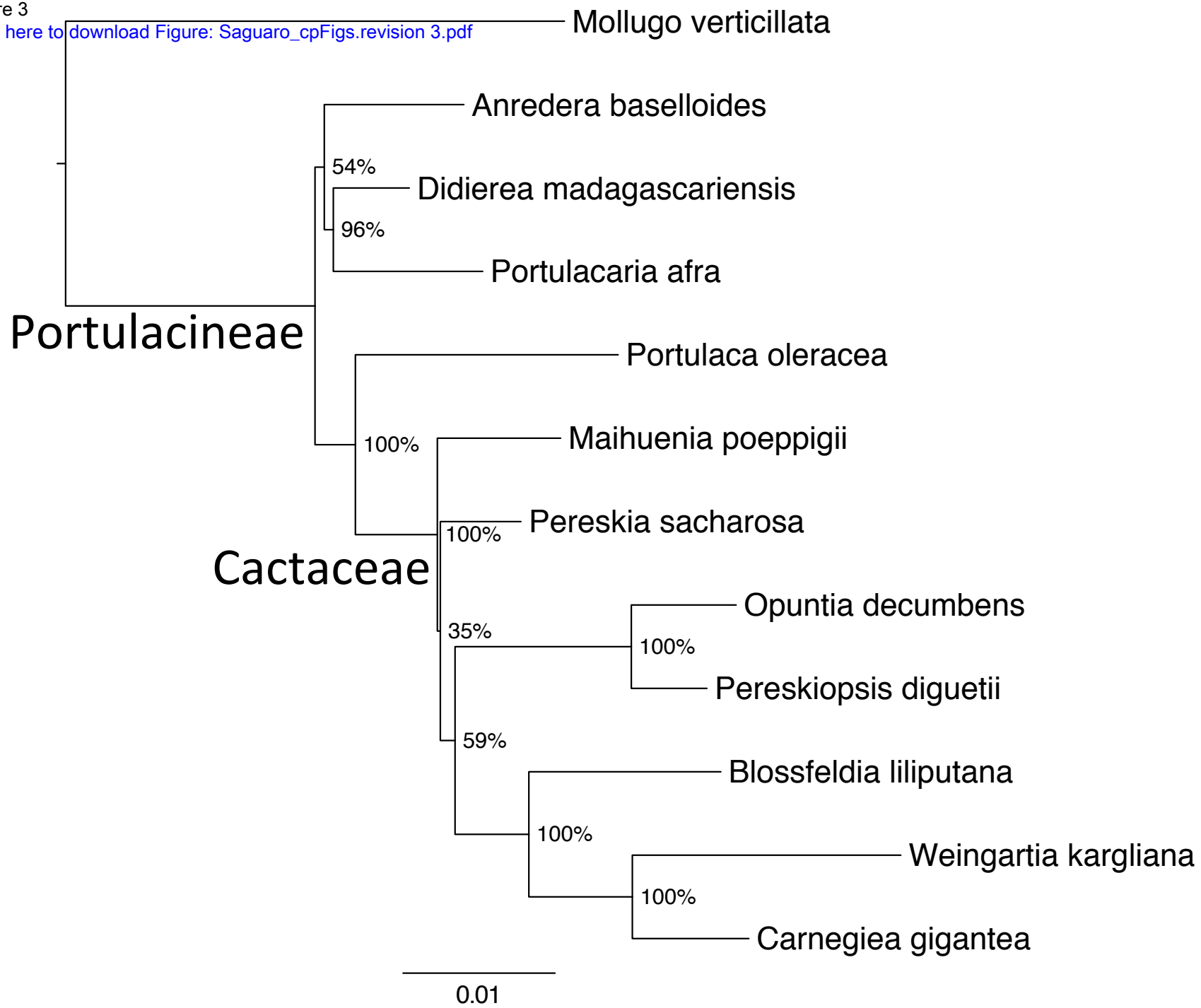

Figure 4

Click here to download Figure: Saguaro_cpFigs.revision 4.pdf

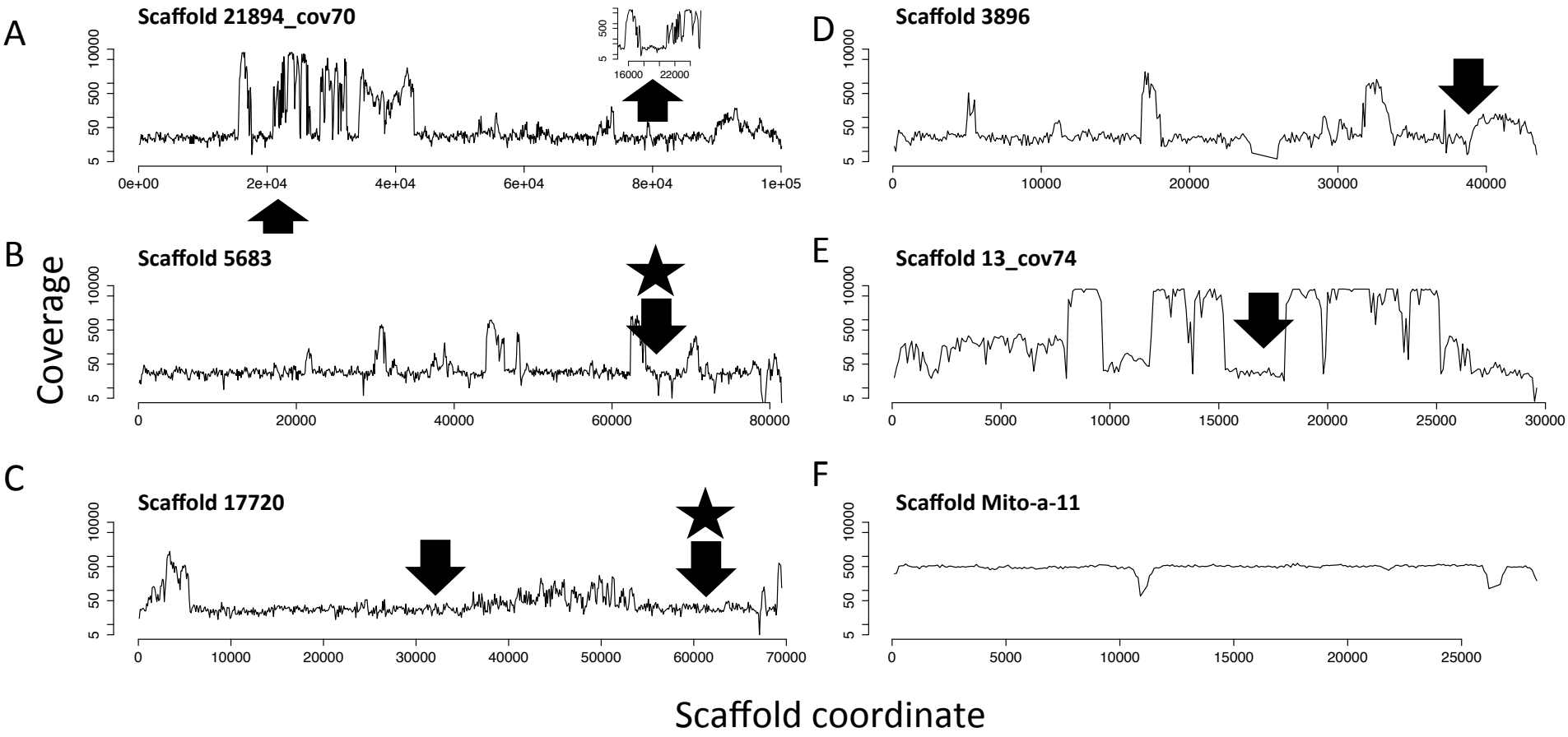

Figure 5

Click here to download Figure: Saguaro_cpFigs.revision 5.pdf